Check for updates

Journal of KIIT. Vol. 23, No. 2, pp. 23-32, Feb. 28, 2025. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2025.23.2.23

수질 예측을 위한 인공지능 학습용 데이터 품질 관리 기법

이상민*[1], 최승호**, 이석훈*[2]

Data Quality Management Methods for AI Training Datasets in Water Quality Prediction

Sangmin Lee*[1], Seungho Choi**, and Sukhoon Lee*[2]

요 약

데이터 산업과 인공지능 시장의 급성장으로 AI 기반 서비스의 신뢰성이 중요해지고 있으며, 이는 데이터 품질과 직결된다. 따라서 데이터 품질 보장을 위해 객관적이고 정량적인 평가 지표가 필요하다. 본 논문은 수질 예측 모델을 위한 인공지능 학습용 데이터셋의 데이터 품질 관리 기법을 제안한다. 이를 위하여 수질 데이터를 수집 및 분석하고, 기존 데이터 품질들을 비교하여 수질 데이터셋의 적합한 데이터 품질 기준을 선정한다. 이후 선정된 데이터 품질 기준에 따라 품질을 변화시키며, 이에 따른 예측 모델의 성능 결과를 분석하여 평가한다. SimpleRNN, LSTM, GRU 세 가지 모델을 통해 각 품질 요소가 모델 성능에 미치는 영향을 분석한다. 연구 결과, 수질 예측 모델에 적합한 정량적 품질 지표를 도출하였다.

Abstract

With the rapid growth of the data industry and artificial intelligence market, the reliability of AI-based services is becoming important, which is directly connected to data quality. Therefore, objective and quantitative evaluation indicators are needed to ensure data quality. This paper proposes a data quality management technique of an AI learning dataset for a water quality prediction model. To this end, water quality data is collected and analyzed, and existing data quality is compared to select appropriate data quality criteria of the dataset. After that, the quality is changed according to the selected data quality criteria, and the performance results of the prediction model are analyzed and evaluated. The effect of each quality factor on the model performance is analyzed through three models: Simple RNN, LSTM, and GRU. As a result of the study, a quantitative quality index suitable for the water quality prediction model was derived.

Keywords
data quality, water guality prediction, quality criteria, artificial intelligence

* 국립군산대학교 소프트웨어학과(*2 교신저자)
- ORCID[1]: https://orcid.org/0009-0003-8907-2923
- ORCID[2]: https://orcid.org/0000-0002-3390-5602
** 국립군산대학교 소프트웨어융합공학과
- ORCID: https://orcid.org/0009-0001-1299-5020

· Received: Nov. 14, 2024, Revised: Dec. 16, 2024, Accepted: Dec. 19, 2024
· Corresponding Author: Sukhoon Lee
Dept. of Software Science & Engineering, Kunsan National University, Korea
Tel.: +82-63-469-8914, Email: leha82@kunsan.ac.kr

Ⅰ. 서 론

최근 데이터는 국내외 산업에서 수집, 정제, 서비스 활용 등 다양한 형태로 활용되고 있으며, 데이터 산업 시장과 더불어 인공지능 분야의 발전으로 인공지능 시장의 규모도 지속적으로 커지고 있다[1][2]. 이와 함께 데이터의 품질 관리 필요성 또한 중요한 요소가 되고 있으며, 데이터 품질 평가를 위해 국내외 다양한 기관들은 데이터 품질 평가 기준을 수립하고 있다. 이는 데이터가 단순한 정보의 집합을 넘어 의사결정과 가치 창출의 핵심적인 요소로 자리 잡았음을 의미한다[3].

수질 오염은 생태계 파괴, 수자원 부족, 공중 보건 문제 등 다양한 환경적, 사회적 문제를 초래하는 중요한 이슈로 자리하고 있다[4][5].

특히, 산업화와 도시화가 가속화되면서 수질 오염 문제는 점차 심화되고 있으며, 이로 인한 사회적, 경제적 손실 역시 증가하는 추세이다. 따라서 수질 상태를 지속적으로 모니터링하고 변화 양상을 예측하는 것은 효과적인 수질 관리를 위해 필수적이다. 현재 다양한 수질 데이터가 실시간으로 수집되고 있으며, 이러한 방대한 수질 데이터는 수질 관리와 오염 예방을 위한 핵심 자원으로써 활용되고 있다[6][7].

수질 데이터 또한 인공지능 기술의 발전으로 수질 예측을 위한 다양한 모델들이 활발히 활용되고 있다. 특히 시계열 데이터 처리에 강점을 보이는 Linear, SimpleRNN, LSTM, GRU, CNN 등이 수질 예측에 적용되고 있다[8]. 이러한 모델들은 과거의 수질 데이터를 학습하여 미래의 수질 상태를 예측하며, 각 모델은 그 특성에 따라 단기 또는 장기 예측에 적합한 성능을 보인다.

현재 수질 예측 모델 연구들은 대부분 알고리즘 개선이나 모델 구조 최적화에 초점을 맞추고 있다. 이러한 인공지능 모델의 성능은 입력 데이터의 품질에도 영향을 받는다. 하지만 학습용 데이터의 품질이 모델 성능에 미치는 영향에 대한 체계적인 연구는 부족한 실정이다. 특히 수질 데이터는 센서 측정의 특성상 정밀도 문제, 결측치 발생, 이상치 존재 등 다양한 품질 관련 문제가 존재하지만, 이러한 데이터 품질 문제가 예측 모델의 성능에 어떤 영향을 미치는지에 대한 정량적인 분석이 미흡하다. 또한 수질 데이터의 특성을 고려한 품질 관리 기준이 명확히 정립되어 있지 않아, 데이터 수집 및 전처리 과정에서 일관된 품질 관리가 어려운 상황이다. 이는 수질 예측 모델의 신뢰성과 정확성을 저하시키는 주요 원인이 될 수 있다.

본 논문은 수질 예측을 위한 인공지능 학습용 데이터의 품질 평가 기준을 수립하고 이를 통한 데이터 품질 관리 기법을 제안한다. 먼저, 수질 데이터를 수집하고 수질 데이터에 적합한 품질 기준을 선정한다. 이러한 품질 요소들이 다양한 수질 예측 모델의 성능에 미치는 영향을 분석하고자 한다. 이를 위하여 수질 예측을 위한 데이터셋으로 AI Hub에서 제공하는 수질 측정 및 오염원 데이터를 활용한다[9]. 본 논문은 실험을 위하여 수질 데이터셋의 특성에 적합한 품질 기준 요소를 정의하고 각 요소가 딥러닝 모델 성능에 어떤 영향을 미치는지 실험적으로 검증한다. 이를 통해 수질 데이터의 효과적인 품질 관리 방안을 제시하고, 나아가 수질 예측 모델의 정확도와 신뢰성을 향상시키기 위한 구체적인 데이터 품질 기준을 수립하고자 한다.

결과적으로 본 논문은 데이터 품질 관리 기준을 수립함으로써 데이터의 신뢰성과 일관성을 높이는 데 기여하고, 향후 수질 데이터의 수집 및 예측 시 데이터 품질 관리에 활용될 수 있다.

본 논문의 구성은 다음과 같다. 제2장에서 데이터 품질 기준 관련 연구와 수질 예측 관련 연구를 소개하고, 제3장에서 수질 데이터셋을 분석하고 데이터 품질 기준을 조사한다. 수질 데이터에 적합한 품질 기준과 실험 방법을 정의한다. 제4장에서 선정된 수질 데이터 품질에 따른 수질 예측 모델 성능을 비교하고 데이터 품질 기준별 정량 지표를 수립한다. 마지막으로 제5장에서 결론을 기술한다.

Ⅱ. 관련 연구

이 장에서는 기존에 활용되고 있는 수질 예측 관련 연구와 데이터 품질 기준을 기술한다[10]-[15].

한국지능정보사회진흥원에서 배포한 “인공지능 데이터 구축&활용 가이드라인 - 환경오염 AI 데이터”은 수질 예측을 위한 AI 모델 개발에 필요한 데

이터 수집, 정제, 검수, 활용 등의 데이터 구축 전반의 지침과 모델 학습 및 활용까지 제공한다[10].

미국농업생물공학회는 논문 “수문학 및 수질 모형 : 성능 측정 및 평가 기준”을 통하여 수문학에서 활용되는 수질 모델의 성능을 평가하는 기준을 제시하고, 수질 모델의 성능을 객관적으로 평가할 수 있는 기준을 제공한다[11]. 이 기준은 하천 유량, 퇴적물, 영양분 등 다양한 수질 관련 모델에 적용할 수 있다. 본 논문은 미국농업생물공학회에서 제시하는 성능 평가 기준을 참조하여 수질 데이터의 품질 관리 기준으로 활용한다.

한국지능정보사회진흥원의 인공지능 학습용 데이터 품질관리 가이드라인은 다양한 분야에서 인공지능이 활용되고 있어 인공지능 학습용 데이터 품질 관리를 체계적으로 수행하기 위해 제정되었다[12]. 데이터 수집, 전처리, 가공, 검수 등 학습용 데이터가 적정한 수준의 품질을 유지하고 있는지 데이터 품질 기준과 품질 관리 지침도 제시하고 있다.

ISO/IEC 25024는 국제 표준화 기구(ISO)와 국제 전기기술 위원회(IEC)가 공동으로 개발한 SQuaRE (Software product Quality Requirements and Evaluation) 프로젝트의 일부로, 데이터 품질 측정에 관한 국제 표준이다[13]. 이 표준은 데이터 품질 특성을 정의하고 이를 측정하기 위한 방법을 제시하고 있다. 정확성, 완전성, 일관성 등 15가지 주요 데이터 품질 특성을 정의하며, 각 특성에 대한 측정 방법과 지표를 제공한다.

한국지능정보사회진흥원의 공공데이터 품질관리 매뉴얼은 공공데이터의 품질 확보 방안을 마련하기 위해 제정되었다[14]. 데이터 품질 관리에 대해 설명하고 있고 서류 위주의 평가를 진행한다. 국내 공공데이터의 체계적인 품질 관리를 위한 지침을 제공하고 공공데이터의 특성을 고려하여 데이터 품질 관리의 기준, 절차, 방법론을 제시하고 있다.

한국데이터베이스진흥원의 데이터 품질진단 절차 및 기법은 증가 추세에 있는 데이터베이스의 품질 진단에 적용하여 도움을 주기 위해 제정되었다[15]. 데이터 품질관리를 위한 체계적인 접근 방법을 제시하고 있으며 데이터 품질진단의 전체 프로세스를 상세히 설명하고, 각 단계별로 적용할 수 있는 구체적인 기법들을 제공한다.

III. 수질 데이터 수집 및 분석

3.1 문제 정의 및 제안 프로세스

수질 예측을 위한 인공지능 모델 개발에 있어 두 가지 핵심적인 문제가 존재한다. 첫째, 모델 학습에 사용되는 수질 데이터의 품질 기준이 명확하지 않다는 점이다. 즉, 어느 수준의 데이터 품질이 확보되어야 신뢰할 수 있는 예측 결과를 얻을 수 있는지에 대한 기준이 부재하다. 둘째, 수질 데이터의 품질을 평가할 수 있는 정량적 평가 지표가 부족하다는 점이다. 수질 데이터의 특성을 고려한 객관적이고 정량적인 품질 평가 방법이 필요하다.

그림 1은 문제를 해결하기 위한 본 논문의 프로세스이다. 먼저 인공지능 학습용 데이터셋을 수집한다. 데이터 분석 단계에서는 수집된 데이터 특성을 파악하기 위해 프로파일링을 진행한다. 품질 기준 비교 및 품질 기준 선정 단계에서는 기존의 품질 기준들을 조사하여 학습할 수질 데이터셋에 적합한 기준을 선정한다. 이어서, 데이터 품질 변화에 따른 인공지능 모델 성능 비교 단계에서는 선정된 품질 기준에 따른 실험 방법을 정의하고 선정된 품질 기준을 변화시켜 인공지능 모델 성능을 비교한다. 마지막으로 정량지표 설정 단계에서는 성능 비교 결과를 통해 품질 변화에 따른 정량 지표를 설정한다.

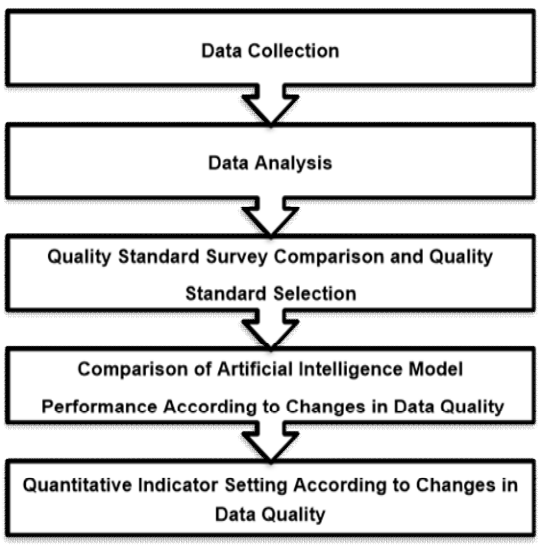

그림 1. AI 데이터 품질 평가 프로세스
Fig. 1. AI data quality assessment process

3.2 수질 데이터셋 수집

본 논문은 사용된 데이터는 AI Hub에서 제공하는 "수질 측정 및 오염원 데이터"를 활용한다. 해당 데이터셋은 물 환경측정망, 수질 자동측정망, 수질 원격감시체계(TMS), 하천 유량 측정, 강수량 등의 데이터를 포함하고 있으며, 본 논문은 이 중 낙동강 권역의 물 환경측정망 자동측정망 데이터를 활용한다[9].

낙동강 권역의 24개 측정소에서 수집된 자동측정망 데이터는 수온, PH, DO, 전기전도도, TOC, 탁도, 클로로필-a, TN, TP 등을 측정한 데이터로 구성되어 있으며, 연구에서는 2013년부터 2020년까지의 데이터를 사용한다. 자동측정망 데이터는 1시간 단위로 측정되며, 총 59개의 수질 관련 변수를 포함하고 있다. 각 변수당 약 13만 건의 데이터가 존재하며, 이 중 8개 항목(수온, PH, DO, 전기전도도, TOC, 클로로필-a, TN, TP)을 분석에 사용한다.

수질 데이터의 정제는 다음과 같은 순서로 진행된다. 1차적으로 물환경측적망 운영계획 관련 법령에 근거하여 1차적으로 데이터가 정제되고 2차적으로 중위절대편차를 사용해 신뢰구간을 이용한 이상치 판단 기법으로 95%의 신뢰도에 따라 신뢰구간 안에 포함되지 않으면 이상치로 제거한다[4].

구축된 수질 데이터셋의 컬럼은 tmpr_value, ph_value, do_value, ec_value, toc_value, 총질소_값, 총인_값, 클로로필-a로 구성되어 있으며, 정의와 특징은 표 1과 같다.

3.3 데이터 프로파일링

데이터 프로파일링은 통계적 분석을 사용하여 데이터셋의 데이터 현상을 파악하여 데이터의 잠재적 오류를 발견할 수 있는 기법이다. 데이터 프로파일링을 위해 데이터의 형식, 데이터의 개수, 결측값의 개수, 최솟값, 최댓값, 평균값, 소수점 자릿수를 측정한다[16][17].

데이터 형식을 통해 각 컬럼이 어떤 유형의 데이터인지 파악할 수 있다. 데이터 건수를 통해 데이터의 크기를 파악하고, 결측치를 파악할 수 있다. 최솟값과 최댓값을 통해 데이터의 범위를 확인할 수 있고 평균값을 통해 데이터의 중심 경향을 파악할 수 있다.

표 2는 구축된 데이터셋의 프로파일링을 한 결과를 보이며 이를 통해 데이터를 파악할 수 있다.

표 1. 수질 데이터셋의 정의와 특징
Table 1. Definition and characteristics of water quality datasets

column	definition	features
tmpr_value	water temperature at the collection site	real number ≥ 0, unit (℃)
ph_value	hydrogen ion concentration at the collection site	real number ≥ 0
do_value	dissolved oxygen at the collection site	real number ≥ 0, unit (mg/L)
ec_value	electrical conductivity at the collection site	real number ≥ 0, unit (mg/L)
toc_value	total organic carbon at the collection site	real number ≥ 0, unit (mg/L)
tn_value	total nitrogen at the collection site	real number ≥ 0, unit (mg/L)
tp_value	total phosphorus at the collection site	real number ≥ 0, unit (mg/L)
chl-a_value	chlorophyll-a at the collection site	real number ≥ 0, unit (mg/L)

표 2. 수질 데이터셋 프로파일링 결과
Table 2. Water quality dataset profiling results

Column	Data types	Missing values	Min values	Max values	Mean values
tmpr_value	float64	0	0.0	35.983	15.947
ph_value	float64	0	5.9	10.19449	7.937
do_value	float64	0	2.00833	17.39167	10.121
ec_value	float64	0	69.80833	448.25833	225.752
toc_value	float64	0	0.23799	25.32833	2.787
tn_value	float64	0	0.45667	10.96283	2.467
tp_value	float64	0	0.0	2.766	0.017
chl-a_value	float64	0	0.0	163.99167	15.236

3.4 수질 데이터 품질 기준 선정

데이터 품질 평가를 진행하기 위해서는 품질 기준을 정하는 것이 필요하다. 이를 위해 본 논문은 국제 표준 및 기관들의 가이드라인을 참고하여, 데이터 품질 기준들을 분석한다[10]-[13].

한국지능정보사회진흥원의 인공지능 학습용 데이터 품질 관리 가이드라인에서는 품질 기준으로 준비성, 완전성, 유용성, 적합성, 정확성, 유효성의 기준을 사용한다[10].

ISO/IEC 25024는 데이터 품질 평가 기준으로 정확성, 완전성, 일관성, 신뢰성, 최신성, 접근성, 준수성, 기밀성, 효율성, 정밀성, 추적성, 이해가능성, 가용성, 이식성, 복구성 등의 기준을 사용한다[11].

한국지능정보사회진흥원의 공공데이터 품질관리 매뉴얼에서는 데이터 품질 기준으로 준비성, 완전성, 일관성, 정확성, 보안성, 적시성, 유용성의 품질 기준을 사용한다[12].

한국데이터베이스진흥원의 데이터 품질진단 절차 및 기법에서는 데이터 품질 평가 기준으로 완전성, 유효성, 정확성, 유일성, 일관성의 품질 기준을 사용한다[13].

본 논문은 앞서 언급한 데이터 품질 기준들을 분석한 결과, 수질 데이터셋의 특성을 고려하여 정량적 품질 평가의 지표를 측정할 수 있는 기준으로 충분성, 완전성, 정밀성, 정확성을 선정한다.

3.5 데이터 품질 기준별 실험 방법

표 3은 선정된 데이터 품질 기준의 정의와 실험 방법이다. 충분성 실험은 학습 데이터의 수량을 2,000건에서 128,728건까지 단계적으로 증가시키며 모델의 성능을 측정한다. 완전성 실험은 데이터의 결측값을 선형 보간법으로 처리하고, 결측 비율을 3%에서 30%까지 변화시키며 모델의 성능을 측정한다. 정밀성 실험는 데이터 수집 간격을 1시간에서 7시간까지 늘려가며 모델의 성능을 측정하며, 정확성 실험은 학습 데이터에 최대 10%까지의 노이즈를 인위적으로 추가하여 모델의 성능을 측정한다. 이러한 실험 방법을 통해 각 품질 기준이 모델 성능에 미치는 영향을 정량적으로 분석하고자 한다.

3.6 데이터 품질의 정량 지표 설정 방법

모델 성능 평가는 수문학 및 환경 과학 분야에서 널리 사용되는 지표인 NSE(Nash-Sutcliffe Effieincy)를 활용한다. NSE는 -∞에서 1 사이의 값을 가지며, 성능이 0보다 작다면 모델의 예측값이 관측값의 평균값으로 대체하는 것보다 못함을 의미하며, 1에 가까울수록 모델의 예측값이 관측값을 정확히 예측한다는 의미이다.

식 (1)은 NSE 계산식을 나타낸다. 각각 O_i는 관측값, P_i는 예측값, O_j 관측값 평균, N은 데이터의 총 개수이다.

표 4는 수질 예측 모델 성능 평가 기준으로 미국농업생물공학회에서 정의한 수질 모델 성능 평가 기준을 사용한다[14].

$$NSE = 1 - \frac{\sum_{i=1}^{N}(O_i - P_i)^2}{\sum_{i=1}^{N}(O_i - O_j)} \quad (1)$$

표 3. 데이터 품질 기준 선정 및 실험 방법
Table 3. Selection of data quality standards and experimental methods

criterion	definition	experimental method
sufficiency	criterion for evaluating the quantitative aspect of data required for AI model training	model performance comparison by increasing data quantity from 2,000 to 128,728 to determine suitable training volume
completeness	criterion for evaluating the handling of missing values in training data	model performance comparison by using linear interpolation to process missing values and increasing the missing ratio from 3% to 30%
precision	criterion for evaluating the sensor data collection frequency in training data	model performance comparison by extending data collection intervals from 1 hour to 7 hours
accuracy	criterion for evaluating the extent of noise in training data	model performance comparison by artificially adding noise from 0% to 10% in the training data

표 4. 성능 평가 기준
Table 4. Performance evaluation criteria

Performance	Criteria
Very good	NSE > 0.8
Good	0.7 < NSE ≤ 0.8
Satisfactory	0.5 < NSE ≤ 0.7
Not satisfactory	NSE ≤ 0.5

IV. 실험 및 결과

4.1 실험 방법

본 논문에서는 데이터 품질에 따른 수질 예측 모델 성능 비교를 위해 순환 신경망 모델 중 SimpleRNN, LSTM, GRU를 사용한다. 이 모델들은 시계열 데이터 처리에 강점을 가지고 있어 수질 데이터와 같은 시간 의존적 데이터 분석에 적합하다.

실험에 사용된 데이터셋은 2013년부터 2020년까지의 수질 측정 데이터로 구성되어 있다. 입력 데이터로는 수온, pH, DO(용존산소량), 전기전도도, TOC(총유기탄소), 총질소, 총인, 클로로필-a의 8개 항목을 사용하며, 항목별로 128,728건을 사용한다. 출력 데이터로는 수질 상태를 대표하는 5개 항목(DO, TOC, 총질소, 총인, 클로로필-a)을 사용한다. 본 실험에서는 3일간의 시계열 데이터를 입력으로 사용하여 3일 후의 수질 상태를 예측한다.

데이터 품질이 예측 모델의 성능에 미치는 영향을 분석하기 위해 충분성, 완전성, 정확성, 정밀성의 네 가지 품질 기준을 실험 대상으로 선정한다. 각 품질 기준에 대해 SimpleRNN, LSTM, GRU 세 가지 모델의 성능 변화를 측정하고 비교 분석한다.

4.2 충분성 실험

충분성은 인공지능 모델 학습에 필요한 적정 데이터 양을 의미한다. 이를 평가하기 위해 데이터 수량을 2,000개부터 128,728개까지 증가시키며 모델 성능을 비교한다. 데이터 수량에 따른 모델의 예측 성능 변화를 분석한다. 이를 통해 예측 성능을 보장하는 최소한의 데이터 수량을 파악할 수 있다.

그림 2 (a), (b), (c)은 데이터 수량 변화에 따른 SimpleRNN, LSTM, GRU 모델의 예측 성능 변화를 보여준다. 세 모델 모두 데이터 수량이 증가함에 따라 전반적인 성능 향상을 보인다. 특히 20,000건 이상에서는 Satisfactory 수준의 안정적인 예측 성능을 보이기 시작했으며, 60,000건 이상에서는 대부분 Very Good 수준의 우수한 예측 성능을 달성했다.

4.3 완전성 실험

완전성은 결측치 없이 데이터가 얼마나 완전한지를 나타낸다. 실험을 위해 인위적으로 결측치를 생성하고, 이를 선형 보간법으로 처리한다. 결측 비율을 3~30%까지 3%의 간격으로 증가시키며 모델 성능 변화를 분석한다. 이를 통해 결측치를 선형 보간법으로 처리했을 때 모델 성능에 미치는 영향을 파악할 수 있다.

그림 2 (d), (e), (f)는 결측치 비율 변화에 다른 SimpleRNN, LSTM, GRU 모델의 예측 성능 변화를 보여준다. 분석 결과, 세 모델 모두 결측치 비율이 증가하더라도 Very Good 수준의 예측 성능을 유지했다. 이는 선형 보간법을 통한 결측치 처리가 효과적임을 보여준다.

만약 보간법을 사용하지 않는 경우에는 해당 결측치들을 삭제해야 하며, 이러한 경우 결측치의 비율에 따라 충분성에 영향이 미칠 수 있다. 따라서, 식 (2)는 결측치 비율에 따라 삭제 후 남는 유효데이터 수를 의미하며, 결측치 비율이 커질수록 학습을 위한 유효데이터 수는 줄어든다.

$$n_{valid} = n_{train} \times (1 - r_{missing}) \quad (2)$$

각각 n_{valid}은 유효데이터 수, n_{train}은 학습 데이터 수, $r_{missing}$은 결측치 비율이다.

예를 들어, 학습에 필요한 최소 데이터 수가 12,000건일 때 15,000건의 데이터를 수집하더라도 결측치 비율이 20%라면 실제 학습에 사용할 수 있는 데이터는 12,000건이 남으므로 충분성으로 평가 시 영향을 줄 수 있다.

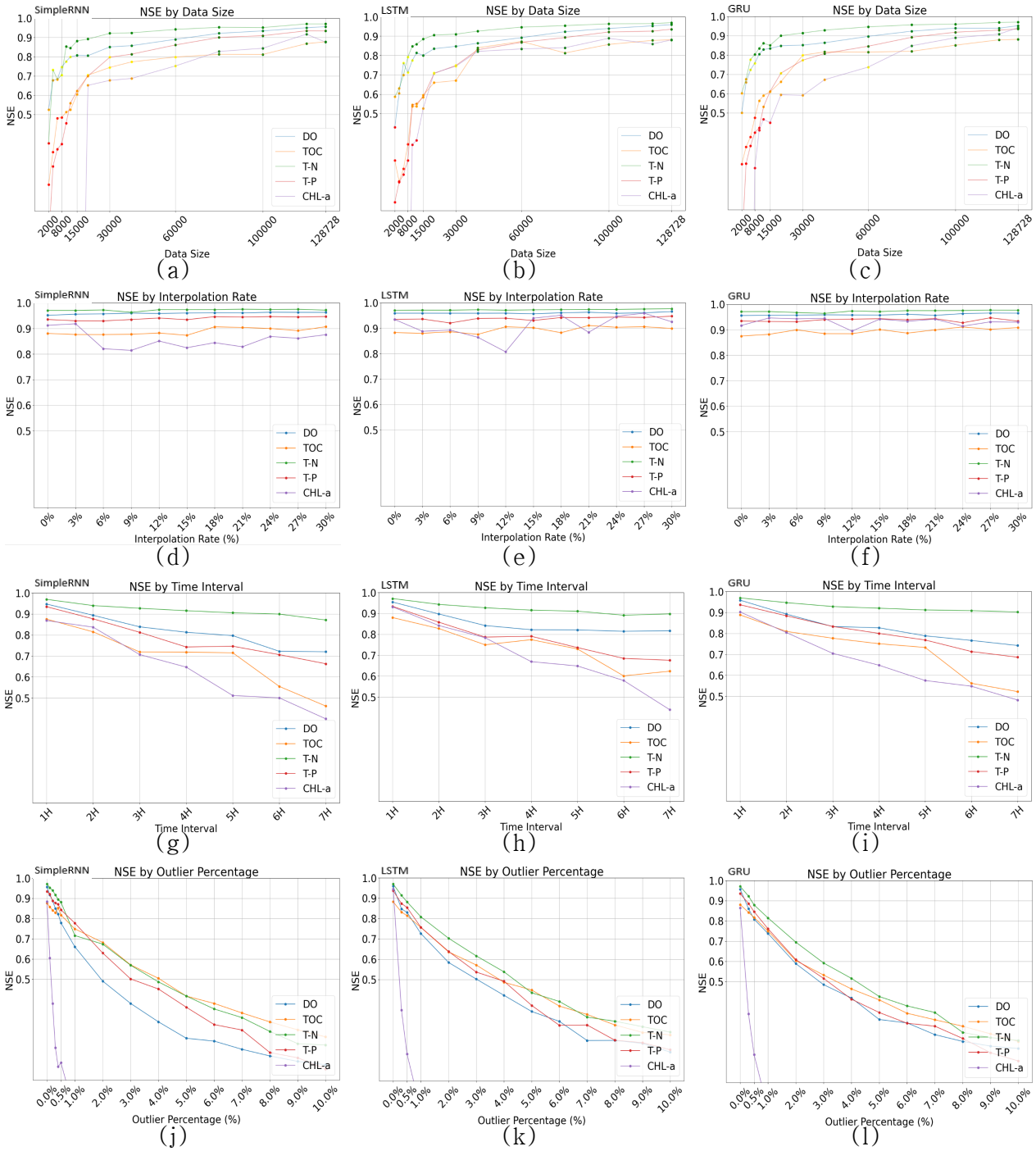

그림 2. 데이터 품질 기준에 따른 예측 모델 성능 결과

Fig. 2. Predictive model performance results based on data quality criteria

4.4 정밀성 실험

정밀성은 데이터 수집 간격의 적절성을 의미한다. 이를 평가하기 위해 데이터 수집 간격을 1시간부터 7시간까지 1시간 단위로 증가시키며 모델 성능을 분석한다. 데이터 수집 간격 변화에 따른 모델의 성능 변화를 분석한다. 이를 통해 수질 예측에 필요한 적절한 데이터 수집 간격을 파악할 수 있다. 그림 2 (g), (h), (i)은 데이터 수집 시간 간격 변화에 따른 SimpleRNN, LSTM, GRU 세 모델의 예측 성능 변화를 보여준다. 세 모델 모두 시간 간격이 증가함에 따라 예측 성능이 저하되고 2시간 간격에

서는 모든 모델이 Very Good 수준의 성능을, 3~5시간 간격에서는 대부분 Good 수준 마지막으로 6시간 간격에서는 Satisfactory 수준의 성능을 만족한다.

4.5 정확성

정확성은 데이터 값의 정확도를 의미한다. 본 실험에서는 학습 데이터셋에 원본 데이터의 ±10%의 노이즈를 추가하여 데이터셋의 정확성을 저하시켜 데이터에 0~10%까지 증가시키며 모델 성능 변화를 분석한다. 이를 통해 데이터의 정확성 저하가 모델 성능에 미치는 영향을 파악할 수 있다.

그림 2 (j), (k), (l)은 노이즈 비율 변화(0~10%)에 따른 SimpleRNN, LSTM, GRU 모델의 예측 성능 변화를 보여준다. 분석 결과, 세 모델 모두 노이즈 비율이 증가함에 따라 예측 성능이 급격히 저하되는 경향을 보였으며. 특히 2~4% 이상의 노이즈에서는 Not Satisfactory 수준 이하로 성능이 크게 감소했다.

4.6 정량 지표 설정 결과

표 5는 SimpleRNN, LSTM, GRU 세 가지 모델에 대한 데이터 품질 기준별 정량 지표를 보여준다. 이 지표는 모델별 예측 성능에 미치는 각 품질 요소의 영향을 평가하기 위한 기준이다.

표 6은 최종적으로 선정된 정량 지표 설정 결과이며 본 논문에서는 데이터 품질을 최적화하기 위해 세 모델의 데이터 품질 기준을 분석하여 각 평가 기준별로 가장 높은 수준의 기준값을 적용한다.

하지만, 클로로필-a 변수의 경우, 품질에 따른 성능 차이가 크게 나타나 지표 설정이 어려운 변수로 분석되어 본 논문의 정량 지표 설정에서 제외한다.

Ⅴ. 결 론

본 논문에서는 수질 데이터의 품질이 인공지능 예측 모델의 성능에 미치는 영향을 체계적으로 분석한다. 데이터 품질 기준을 선정하고, 이에 따른 SimpleRNN, LSTM, GRU 모델의 성능 변화를 다섯 가지 수질 지표에 대해 평가한다. 모델의 성능 평가에는 NSE 지표를 사용하며, 성능 평가 기준을 적용하여 성능을 분류한다.

표 5. 모델별 데이터 품질 기준에 따른 정량 지표
Table 5. Quantitative indicators according to data quality standards for each model

		Very good NSE>=80	Good NSE>=70	Satisfactory 70>NSE>=50	Not satisfactory NSE<50
SimpleRNN	Sufficiency	≥80,000 records	≥30,000 records	≥12,000 records	<12,000 records
	Precision	≤ 2H	≤ 5H	≤ 6H	≥7H
	Accuracy	≤ 0.3%	≤ 0.5%	≤ 1%	≥2%
LSTM	Sufficiency	≥40,000 records	≥30,000 records	≥10,000 records	<10,000 records
	Precision	≤ 2H	≤ 3H	≤ 6H	≥7H
	Accuracy	≤ 0.5%	≤ 1%	≤ 3%	≥4%
GRU	Sufficiency	≥40,000 records	≥30,000 records	≥12,000 records	<12,000 records
	Precision	≤ 2H	≤ 5H	≤ 7H	>7H
	Accuracy	≤ 0.5%	≤ 1%	≤ 2%	≥3%

표 6. 최종 선정된 데이터 품질 기준 정량 지표
Table 6. Finally selected quantitative indicators of data quality standards

	Very good	Good	Satisfactory	Not satisfactory
Sufficiency	≥ 80,000 records	≥ 30,000 records	≥ 12,000 records	< 12,000 records
Completeness	Evaluated based on sufficiency after applying formula (2)			
Precision	≤ 2H	≤ 3H	≤ 6H	>7H
Accuracy	≤ 0.3%	≤ 0.5%	≤ 1%	≥2%

본 연구를 통하여 수질 데이터의 특성을 고려한 품질 기준을 제시하고, 데이터 품질과 인공지능 모델 성능 간의 상관관계를 정량적으로 분석함으로써, 수질 예측이나 수질 관리 시스템의 개선에 기여할 수 있을 것으로 기대된다. 더불어 다른 환경 데이터나 시계열 데이터에도 적용될 수 있는 확장성을 가질 것으로 예상된다.

향후에는 본 논문의 방법론을 다양한 환경 데이터에 확장 적용하여, 환경 데이터의 일반적인 품질 기준을 마련하는 연구를 진행할 예정이다.

References

[1] K. E. Park, J. H. Park, J. Y. Lee, and S. B. Lee, "Method and Application Case of Quality Verification of File Data in Structured Format (CSV) Using ISO/IEC 25024 Quality Characteristics", Proc. of the Korean Institute of Communications and Information Sciences Symposium, pp. 1699-1700, Jun. 2024.

[2] Korea Data Industry Promotion Agency, "2023 Data Industry White Paper", pp. 76-81, Oct. 2023.

[3] S. Park, K. Lee, and A. Lee, "An Empirical Study on the Effects of Source Data Quality on the Usefulness and Utilization of Big Data Analytics Results", Journal of Information Technology Applications & Management, Vol. 24, No. 4, pp. 197-214, Dec. 2017. https://doi.org/10.21219/jitam.2017.24.4.197.

[4] C.-G. Park, D.-I. Jo, G.-Y. Choi, J.-R. Song, J.-W. Lee, and J.-Y. Park, "Efficient Water Resource Management in Response to Climate Change: An Analysis of the Economic Impact of Water Pollution Improvement Projects", The Studies in Regional Development, Vol. 51, No. 2, pp. 81-108, Dec. 2019. http://doi.org/10.35526/srd.2019.51.2.005.

[5] M.-J. Choi, W.-H. Jung, J.-H. Choi, and Y.-I. Kim, "Study on Water Quality Improvement Based on Pollutant Discharge Characteristics Analysis and Water Quality Modeling in the Seokmun Lake Basin", Journal of Korean Society of Environmental Engineers, Vol. 39, No. 10, pp. 581-590, Oct. 2017. https://doi.org/10.4491/KSEE.2017.39.10.581.

[6] Y.-H. Woo, J.-G. Lee, S.-B. Jang, and Y.-W. Ko, "Design of a Water Pollution Prediction System Using Big Data Analysis", Proc. of the Korea Information Science Society Conference, Busan, Korea, pp. 254-256, Dec. 2017.

[7] B.-S. Gal, J.-B. Park, S.-M. Kim, S.-M. Shin, S.-J. Jang, M.-J. Jeon, and D.-H. Lee, "Study on the Analysis of Target Substances for Management Using Long-term Water Quality Monitoring Data in Tributaries of the Nakdong River Basin", J. Korean Wetlands Soc., Vol. 25, No. 4, pp. 326-334, Nov. 2023. https://doi.org/10.17663/JWR.2023.25.4.326

[8] B.-S. Chea, J.-H. Koo, S.-H. Lee, J.-C. Kwon, S.-H. Kong, and K.-B. Song, "Development of an Algal Bloom Prediction Model Based on Machine Learning", NEAR & Future INSIGHT, Vol. 4, Korea Information Society Development Institute (KISDI), Nov. 2017.

[9] https://www.aihub.or.kr/ [accessed: Oct. 30, 2024]

[10] D. Kim, H. Choi, J. Noh, S. Kim, J. Lee, Y. Kim, and Y. Ko, "Artificial Intelligence Data Construction and Utilization Guidelines v2.4.2 - Environmental Pollution AI Data (Water Quality Measurement and Pollution Source Data)", National Information Society Agency, pp. 1-106, Mar. 2021.

[11] D. N. Moriasi, M. W. Gitau, N. Pai, and P. Daggupati, "Hydrologic and Water Quality Models: Performance Measures and Evaluation Criteria", Transactions of the ASABE (American Society of Agricultural and Biological Engineers), Vol. 58, No. 6, pp. 1763-1785, 2015. https://doi.org/10.13031/trans.58.10715

[12] Y. Ko, J. Park, H. Oh, and H. Yu, "Artificial Intelligence Learning Data Quality Management

Guidelines", National Information Society Agency, pp. 1-205, Feb. 2021.

[13] ISO/IEC, "Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Measurement of data quality", ISO/IEC 25024:2015, pp. 1-45, Oct. 2015.

[14] B. Seo, "Open Government Data Quality Management Manual v2.0", NATIONAL INFORMATION SOCIETY AGENCY, pp. 25-164, Jan. 2018.

[15] C. Lee, S. Kim, S. Sin, J. Seo, S. Lim, D. Lee, S. Park, and J. Myung, "Data Quality Assessment Procedure Manual", Korea Database Agency, pp. 1-156, Oct. 2009.

[16] S. Kim, "A study on patent data quality assessment methods based on data profiling", Proceedings of Korean Institute of Industrial Engineers Conference, Gwangju, Korea, pp. 2008-2011, Apr. 2019.

[17] S. Kim, J. Choi, E. Lee, D. Jeong, and S. Lee, "Data Quality Assessment of DSEM-Traj2018 based on Data Profiling", Proceedings of KIIT Conference, Jeju, Korea, pp. 481-483, Jun. 2021.

저자소개

이 상 민 (Sangmin Lee)

2017년 3월 ~ 현재 : 군산대학교 소프트웨어학과 학부과정

관심 분야 : 데이터 품질, 데이터 분석

최 승 호 (Seungho Choi)

2013년 2월 : 가톨릭관동대학교 경영학과(학사)

2017년 2월 : 가톨릭관동대학교 경영학과(석사)

2024년 2월 ~ 현재: 국립군산대학교 소프트웨어융합공학과 박사과정

관심분야 : 소프트웨어 공학, 딥페이크 탐지, 빅데이터 분석

이 석 훈 (Sukhoon Lee)

2009년 2월 : 고려대학교 전자및정보공학부(학사)

2011년 2월 : 고려대학교 컴퓨터·전파통신공학과(공학석사)

2016년 2월 : 고려대학교 컴퓨터·전파통신공학과(공학박사)

2016년 3월 ~ 2017년 3월 : 아주대학교 의료정보학과 연구강사

2017년 4월 ~ 현재 : 국립군산대학교 소프트웨어학과 부교수

관심 분야 : 사물인터넷, 메타데이터 레지스트리, 데이터 품질, 연합 학습